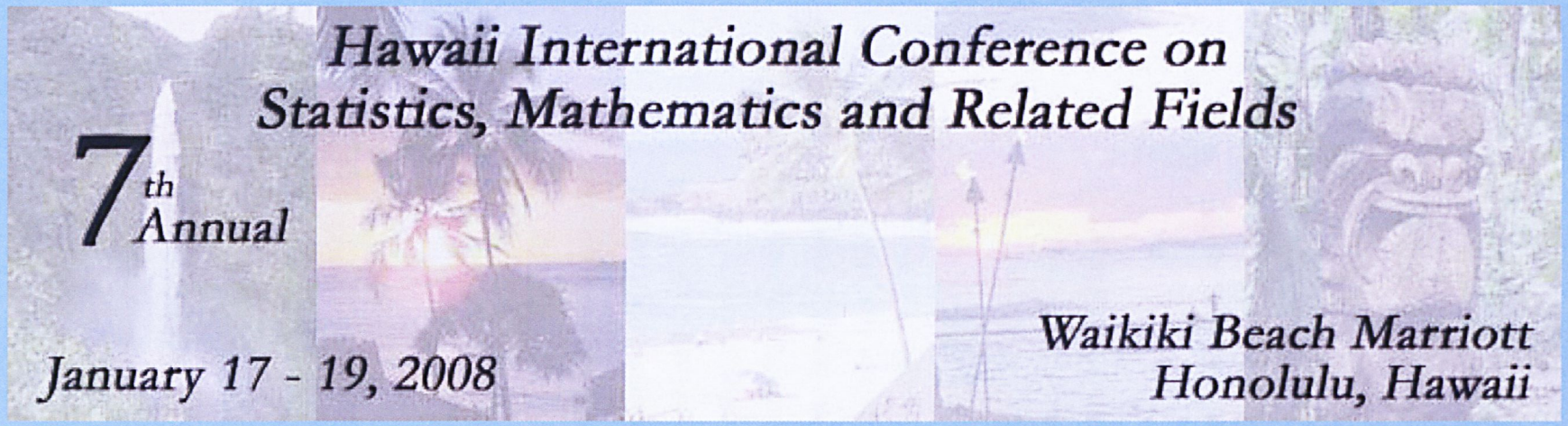
Outliers Labeling With Boxplot Constructed Based on the Shape of Univariate Data Set

By:

Sim, Chiaw Hock

Paper presented at the *7th Hawaii International Conference on Statistics, Mathematics and Related Fields* held on 17-19 January 2008 in Hawaii, USA



2008 Conference Proceedings

We would like to thank all those who attended the 2008 Hawaii International Conference on Statistics, Mathematics and Related Fields. We look forward to seeing you at the 8th Annual Conference to be held in 2009. Please check the website this Spring for dates and further details.

To search for a specific paper presented, or to browse all of the proceedings, please click the appropriate button on the right.

Papers by Topic Area

Papers by Author Name

Browse Proceedings

Hawaii International Conference on Statistics, Mathematics and Related Fields
PO Box 75023, Honolulu, HI 96836
808-542-4081 (phone) / 808-947-2420 (fax)

ISSN #: 1550-3747

Outliers Labeling With Boxplot Constructed Based On The Shape Of Univariate Data Set

C.H. Sim
Institute of Mathematical Sciences
University of Malaya
50603 Kuala Lumpur
Malaysia

Abstract: The graphical boxplot has been widely used in routine screening of outliers in univariate data set. A drawback of the conventional boxplot is that it tends to declare more outliers than those are actually present. It is constructed without taking into consideration the number of observations and the distribution of the data set. To overcome this drawback, a common approach is to determine the lower and upper fences of the boxplot based on the known distribution of the data set and by controling the risk of falsely declares observations in an outlier-free sample of n observations as outliers. However, in practice, the distribution of the data set under study is usually unknown. In this paper, we construct a boxplot with its lower and upper fences evaluated based on the shape of the given data set in terms of its robust measure of skewness and kurtosis. Examples are given to illustrate the proposed outliers detection procedure.

Keywords: Modified boxplot, robust measure of skewness and kurtosis, Tukey's *gh*-distribution, Tukey's *lambda*-distribution.

1 Introduction

An outlier is a member of a subset of observations that appears to be inconsistent with the remainder of a given data set. For a data set taken from an assumed family of univariate statistical distributions, a range may be derived in which the given data are expected to fall within. Data points that fall outside the range may then be identified as potential outliers. There exists an extensive literature on procedures for detecting outliers. Comprehensive discussions on formal outlier detection tests were given by Hawkins (1980) and Barnett and Lewis (1994). These formal tests are mostly restricted to the detection of one or two outliers in the data set with known or assumed distribution. An useful and simple alternative procedure that can be used for a routine screening of multiple outliers is the popular graphical boxplot.

The conventional boxplot of Tukey (1977) declares observations as outliers if they lie beyond the interval

$$[Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)] \tag{1}$$

where k is a fixed constant value, and Q_1 and Q_3 are the first and third quartiles of the data set, respectively. Observations that lie beyond the interval evaluated with $k = 1.5$ are declared as suspected outliers, and those lie beyond the interval evaluated with $k = 3.0$ are extreme outliers. The customarily chosen values of $k = 1.5$ and 3.0 are inappropriate as they are chosen without taking into consideration (i) the number of observations, (ii) the risk of labeling regular observations as outliers, and (iii) the shape of the data set under study. The resulting boxplot tends to declare more outliers than those are actually present in a sample taken either from a skewed distribution, or from a symmetric distribution with long tails. Hoaglin, Mosteller and Tukey (2000) also pointed out that this drawback occurs even when the data set is sampled from a normal distribution. To take into consideration the number of observations in the data set, Hoaglin, Iglewicz and Tukey (1986),

and Hoaglin and Iglewicz (1987) determine the appropriate value of k under the requirement that for an outlier-free sample of n observations taken from a normal distribution, the probability that one or more observations in the sample will be falsely classified as outliers is equal to a given small value α. The idea of using a fixed value α is analogous to the use of a fixed significance level to which users of hypotheses testing are accustomed.

Construction of boxplot to detect outliers in samples taken from known distributions, particularly the normal distribution, has been well documented in the literature (see eg. Davies and Gather, 1993; Carey, Walters, Wager and Rosner, 1997; Sim, Gan and Chan, 2005; and Banerjee and Iglewicz, 2007). They declare observations in a given data set as outliers if they lie beyond the interval

$$[Q_1 - k_l(Q_3 - Q_1), Q_3 + k_u(Q_3 - Q_1)] \tag{2}$$

in which $k_l = k_u$ only if the given data set is taken from a symmetric distribution. The two limits of the above intervals are the lower and upper fences of the boxplot and are general denoted as LF and UF, respectively. The values of k_l and k_u can easily be obtained when the parameters of the known distribution of the data set are known, whereas numerical and simulation approaches are required when the parameters of the given distribution are unknown. These are reviewed and discussed with examples in Sections 2.1 and 2.2.

In practice, the distribution of the data set under study is usually unknown. The main aim of this paper is to determine the values of k_l and k_u by fitting a general distribution to the given data set taken from a population with unknown distribution. Section 3.2 reviews the Tukey's *lambda*-distribution (Freimer, Kollia, Mudholkar and Lin, 1988), and Tukey's *gh*-distribution (Martinez and Iglewicz, 1984) that can be fitted to the shape of the given data set in terms of its robust coefficients of skewness and kurtosis. Determination of k_l and k_u based on the fitted distribution, using an expres-

sion given in Sim et al. (2005), is discussed in Section 3.3. Examples to illustrate the proposed outlier labeling procedure with a boxplot constructed based on the robust coefficients of skewness and kurtosis of the given data set are given in Section 4.

2 Boxplot constructed based on known distribution

2.1 Determination of k_l and k_u with known parameters

The lower and upper quartiles of a random sample $X_1, X_2, \ldots, X_n$ taken from a population, with known distribution function $F_X(x)$, location parameter θ and scale parameter σ, are given as $Q_1 = F_X^{-1}(0.25) = \theta + \sigma F_Z^{-1}(0.25)$ and $Q_3 = F_X^{-1}(0.75) = \theta + \sigma F_Z^{-1}(0.75)$ respectively, where $F_Z(z)$ is the distribution function of the standardized random variable $Z = (X - \theta)/\sigma$. The values of k_l and k_u that result in a probability α of falsely detecting one or more outliers from an outlier-free sample of n observations are then obtained from (2) as

$$k_l = [F_Z^{-1}(0.25) - F_Z^{-1}(\alpha_n/2)]/[F_Z^{-1}(0.75) - F_Z^{-1}(0.25)] \tag{3}$$

and

$$k_u = [F_Z^{-1}(1 - \alpha_n/2) - F_Z^{-1}(0.75)]/[F_Z^{-1}(0.75) - F_Z^{-1}(0.25)] \tag{4}$$

where the value $\alpha_n = P(X < LF) + P(X > UF) = 1 - (1 - \alpha)^{1/n}$ is the error rate that an observation from the outlier-free sample of size n is falsely labeled as outlier. Equations (3) and (4) are derived by taking $P(X < LF) = P(X > UF) = \alpha_n/2$. They are difference from those given in Banerjee and Iglewicz (2007) that are derived by taking $P[min(X_1, \ldots, X_n) < LF] = P[max(X_1, \ldots, X_n) > UF] = \alpha/2)$ which leads to $P(X < LF) = P(X > UF) = 1 - (1 - \alpha/2)^{1/n}$.

Example 2.1:

URL: http://mc.manuscriptcentral.com/lssp E-mail: comstat@univmail.cis.mcmaster.ca

Consider a random sample of size $n = 40$ taken from a normal distribution with known mean $\theta = 8.0$ and standard deviation $\sigma = 2.0$. As $F_Z^{-1}(0.75) = -F_Z^{-1}(0.25)$=0.67449 and $F_Z^{-1}(1-\alpha_n/2) = -F_Z^{-1}(\alpha_n/2)$=3.22, with $\alpha_n = 1-(1-\alpha)^{1/40} = 0.0012815$ when $\alpha = 0.05$, thus we have $k_l = k_u = 1.8871$. Observations in the given normal sample that lie beyond the interval [1.5598, 14.4402], are then labeled as outliers.

Example 2.2:

Consider a random sample of size n taken from an exponential distribution $F_X(x) = 1 - \exp[-(x-\theta)/\sigma]$ with known shift parameter θ and scale parameter σ. By applying the result that the inverse distribution function of a standardized exponential random variable Z is $F_Z^{-1}(u) = -log_e(1-u)$ in equations (3) and (4), we have $k_l = \log_e[(4/3)(1-\alpha_n/2)]/\log_e 3$ and $k_u = -\log_e(2\alpha_n)/\log_e 3$. The interval in (2) are then given by

$$\left[\theta - (\log_e(1-.5\alpha_n))\sigma, \theta + (\log_e 4 - log_e(2\alpha_n))\sigma\right].$$

Thus for a sample of size $n = 40$ taken from an exponential distribution with $\theta = 6.0$ and $\sigma = 2.0$ (i.e. with mean and standard deviation equal to 8.0 and 2.0, respectively), observations that lie beyond the interval [6.0, 20.7057] are labeled as outliers with $\alpha = 0.05$. Note that the value of $\log_e(1-.5\alpha_n)$ is close to zero which leads to the result that $LF = \theta$ even when the sample size n is as small as 10. This explains the reason why we are rarely able to detect outliers from the left tail of an exponentially distributed data.

2.2 Determination of k_l and k_u with unknown parameters

When the parameters of the given distribution $F_X(x)$ are unknown, the first quartile Q_1 and third quartile Q_3 required in (2) are to be estimated from the sample data. They are customarily estimated by the lower fourth $X_{l:n}$ and upper fourth $X_{u:n}$ of a given sample of size n. The lower and upper limits of

the interval given in (2) are then evaluated as

$$LF = X_{l:n} - k_l(X_{u:n} - X_{l:n}) \tag{5}$$

$$UF = X_{u:n} + k_u(X_{u:n} - X_{l:n}). \tag{6}$$

There are many ways of computing the lower and upper fourths. We recommend the fourths to be computed by using the depths $l = i + \frac{1}{2}$ when $f = 0$, and $l = i + 1$ when $f > 0$, where i is the integral part and f is the fractional part of $\frac{n}{4} = i + f$; and $u = n - l + 1$. The advantage of using this definition is that the values l and u are integers when $mod(n, 4)$ equals to either 1, 2 or 3. Even in the case when $mod(n, 4)$ is equal to 0, the resulting non-integer values of l and u have its fractional part both equal to $\frac{1}{2}$.

Hoaglin et al. (1986) discussed a multistep procedure to approximate the values of k_l and k_u from (5) and (6). Their study is applicable only to samples taken from a normal distribution. Sim et al (2005) gave an exact expression that can be routinely used to evaluate the values of k_l and k_u for samples taken from a hypothesized or known distribution of any shape. Their expression takes the form

$$\int_{-\infty}^{\infty}\int_{z_{l:n}}^{\infty}\left\{1 - I_{G_u(y_u)}(n-u,1)\left[1 - I_{G_l(y_l)}(1,l-1)\right]\right\} f_{Z_{l:n},Z_{u:n}}(z_{l:n}, z_{u:n})$$

$$dz_{u:n}dz_{l:n} = \alpha \tag{7}$$

where

(i) $y_l = z_{l:n} - k_l(z_{u:n} - z_{l:n})$ and $y_u = z_{u:n} + k_u(z_{u:n} - z_{l:n})$,

(ii) $Z_{l:n}$ and $Z_{u:n}$ are the lower and upper fourths of the standardized variable of X,

(iii) $f_{Z_{l:n},Z_{u:n}}(z_{l:n}, z_{u:n})$ is the joint probability density function (PDF) of $Z_{l:n}$ and $Z_{u:n}$.

(iv) $G_l(y) = F_Z(y)/F_Z(z_{l:n})$ and $G_u(y) = [F_Z(y) - F_Z(z_{un})]/[1 - F_Z(z_{u:n})]$

(v) $I_p(a,b) = \int_0^p t^{a-1}(1-t)^{b-1}dt/B(a,b)$ is the incomplete beta function.

(vi) α is the risk of falsely detecting one or more outliers from an outlier-free sample.

When the given distribution $F_X(x)$ and thus $F_Z(z)$ is symmetric, the value of $k_l = k_u (= k)$ is obtained by solving (7) numerically. When the distribution is asymmetric, by taking $P(X < LF) = P(X > UF)$, we obtain the value k_l by replacing the term $I_{G_u(y_u)}(n-u, 1)$ in (7) with $1 - I_{G_l(y_l)}(1, l-1)$, and obtain the value k_u by replacing the term $1 - I_{G_l(y_l)}(1, l-1)$ in (7) with $I_{G_u(y_u)}(n-u, 1)$. Note that the values of k_l and k_u do not depend on the location parameter θ and scale parameter σ, however, they do depend on the shape parameter of the asymmetric distribution. Values of k_l and k_u are available in Sim et al. (2005) for samples taken from the normal or exponential distribution.

Example 2.3:

Consider a data set of 20 observations taken from a normal distribution:

2.21 1.84 0.95 0.91 0.36 0.19 0.11 0.10 0.18 0.30
0.43 0.51 0.64 0.67 0.93 1.22 1.35 1.73 *5.80* *12.60*

where the latter two observations (in italic) were originally 0.58 and 1.26, but the decimal points were entered at the wrong place.

For this data set, as $mod(n, 4) = 0$ and $n/4 = 5$, we have the depths $l = 5.5$ and $u = 15.5$ which leads to $X_{l:n} = -0.275$ and $X_{u:n} = 1.075$. Under the assumption that the sample is taken from a normal distribution, the lower and upper fences of a boxplot with $\alpha = 0.05$ are constructed using $k_l = k_u = 2.2382$ (taken from Table 1 of Sim et al., 2005) as $LF = -3.297$ and $UF = 4.097$. Therefore the two large extreme values 5.80 and 12.60 that fall above the upper fence are declared as outliers.

3 Construction of boxplot with unknown distribution

In practice, the distribution of a data set under study is usually unknown. We shall now discuss how to construct a boxplot with its lower and upper fences evaluated by fitting a distribution to the shape of the given data set. The shape of the distribution of a random variable X is commonly measured by the coefficients of skewness and kurtosis which are defined as

$$\sqrt{\beta_1} = E\left[(X-\mu)^3/\sigma^3\right] \text{ and } \beta_2 = E\left[(X-\mu)^4/\sigma^4\right] - 3$$

where $\mu = E(X)$ and $\sigma^2 = E[(X-\mu)^2]$. However, as $\sqrt{\beta_1}$ and β_2 are evaluated using the third and fourth moments of the given data set, they are sensitive to outliers. One or more outliers in either tail of the data set can unduly change the values of $\sqrt{\beta_1}$ and β_2, and the distribution fitted based on these two conventional measures would fail to capture the "true shape" of the data set and would subsequently fail to construct an appropriate boxplot for detecting the outliers. Therefore, in constructing boxplot to detect outliers, it is of the utmost importance that the shape of the given data set should be captured using measures of skewness and kurtosis that are robust to the present of outliers.

3.1 Robust skewness and kurtosis

There are several robust measures of skewness and kurtosis. One of the latest robust measure of skewness is the computational intensive method of "medcouple" introduced by Brys, Hubert and Struyf (2004). However, for ease of estimating the unknown parameters of the fitted distribution from the coefficients of skewness and kurtosis, we adopt the following robust measures of skewness and kurtosis that are evaluated based on the octile of the fitted distribution.

The robust skewness of Hinkley (1975) is defined as

$$SK_q = \frac{[F^{-1}(1-q) - F^{-1}(1/2)] - [(F^{-1}(1/2) - F^{-1}(q)]}{F^{-1}(1-q) - F^{-1}(q)} \tag{8}$$

for any q between 0 and 0.5, where $F^{-1}(q)$ is the qth quantile of the distribution $F(x)$. In this paper, we shall take $q = 0.125$ as recommended in the literature. Moors (1988) proposed a robust kurtosis based on octile that take the form

$$KR = \frac{[F^{-1}(7/8) - F^{-1}(5/8)] + [F^{-1}(3/8) - F^{-1}(1/8)]}{F^{-1}(3/4) - F^{-1}(1/4)}. \tag{9}$$

The bell shape of the normal distribution is characterized by $SK_{0.125} = 0.0$ and $KR = 1.23310$. The highly skewed exponential distribution has $SK_{0.125} = 0.36307$ and $KR = 1.30626$, whereas the symmetric but long-tails Laplace distribution has $SK_{0.125} = 0.0$ and $KR = 1.58495$.

3.2 Two generalized Tukey's univariate distributions

The next step is to fit a general distribution to the underlying distribution of the data set based on the two robust measures of shape given in (8) and (9). We review two well established generalized statistical distributions in which their quantiles and thus their robust coefficients of skewness and kurtosis can easily be obtained. They are the generalized Tukey's *lambda*- and *gh*-distributions.

The generalized Tukey's *Lambda*-distribution

The generalized Tukey's *lambda*-distribution (Freimer et al., 1998) is defined as the distribution of a random variable generated from the transformation

$$X_U = \frac{U^{\lambda_1} - 1}{\lambda_1} - \frac{(1-U)^{\lambda_2} - 1}{\lambda_2} \tag{10}$$

where λ_1 and λ_2 are the shape parameters, and U is a uniform $U(0,1)$ random variable. The random variable X_U is symmetric when $\lambda_1 = \lambda_2$. It follows a logistic distribution when $\lambda_1 = \lambda_2 = 0$, and an exponential distribution

when $\lambda_1 = \infty$, $\lambda_2 = 0$. The normal distribution is closely approximated with $\lambda_1 = \lambda_2 = 0.1349$. The qth quantile $x_U(q)$ of the *lambda*-distribution is obtained by replacing the random variable U in (10) with the value $q \in (0, 1)$, i.e.

$$x_U(q) = \frac{q^{\lambda_1} - 1}{\lambda_1} - \frac{(1-q)^{\lambda_2} - 1}{\lambda_2}.$$

By substituting the $F^{-1}(q)$ by $x_U(q)$ in (8) and (9), we have

$$SK_{0.125}(\lambda_1, \lambda_2) = \frac{\lambda_2 8^{\lambda_2}(7^{\lambda_1} - 2^{\lambda_1} + 1) - \lambda_1 8^{\lambda_1}((7^{\lambda_2} - 2^{\lambda_2} + 1)}{\lambda_2 8^{\lambda_2}(7^{\lambda_1} - 1) + \lambda_1 8^{\lambda_1}(7^{\lambda_2} - 1)} \tag{11}$$

and

$$KR(\lambda_1, \lambda_2) = \frac{\lambda_2 8^{\lambda_2}(7^{\lambda_1} - 5^{\lambda_1} + 3^{\lambda_1} - 1) + \lambda_1 8^{\lambda_1}(7^{\lambda_2} - 5^{\lambda_2} + 3^{\lambda_2} - 1)}{\lambda_2 8^{\lambda_2}(6^{\lambda_1} - 2^{\lambda_1}) + \lambda_1 8^{\lambda_1}(6^{\lambda_2} - 2^{\lambda_2})}. \tag{12}$$

To fit the Tukey's *Lambda*-distribution to the shape of the given data set, we shall then estimate the unknown parameters λ_1 and λ_2 by equating the equations (11) and (12) to their corresponding estimated values and solving for λ_1 and λ_2 numerically.

The generalized Tukey's gh-distribution

The generalized Tukey's gh-distribution (Martinez and Iglewicz, 1984; Hoaglin, 1985) is generated by transforming the standard normal variable Z to

$$X_Z = \frac{e^{gZ} - 1}{g} e^{hZ^2/2} \tag{13}$$

where g and h are the shape parameters. When $g = 0$, the random variable X_Z is symmetric with increasingly heavy tails as h increases, and follows a standard normal distribution when $g = h = 0$. For the case when $g > 0$ and $h = 0$, X_Z follows a lognormal distribution. Distributions which are skewed to the left can be fitted by taking $g < 0$.

For $h > 0$, the qth quantile $x_Z(q)$ of the gh-distribution can be obtained by replacing the random variable Z in (13) with the qth quantile of the standard normal distribution z_q, i.e.

$$x_Z(q) = \frac{e^{gz_q} - 1}{g} e^{hz_q^2/2}. \tag{14}$$

For $h < 0$ and $g > 0$, it was pointed out by Martinez and Iglewicz (1984) that equation (13) can only be used to approximate the quantiles of X_Z when Z falls within the interval $a < Z < b$ over which the function X_Z is monotonic, where a and b are the solutions of the equation $exp(gZ) = hZ/(g + hZ)$. When $h < 0$ and $g = 0$, a and b are the solutions of $Z^2 = -1/h$. For the second data set discussed in Section 4 that has $\hat{h} = -0.30888$ and $\hat{g} = 0.17444$, since $a = -2.2709$, $b = 7.7422$ and the z_q values used to generate the lower and upper fourths $X_{l:n}$ and $X_{l:n}$ of the variable X_Z required in equation (7) lie inside the interval $(-2.2709, 7.7422)$, thus equation (14) can be used to approximate the qth quantile $x_Z(q)$ of X_Z.

By using the results that $z_q = -z_{1-q}$ and $z_{0.5} = 0$, the robust skewness and kurtosis of the gh-distribution are given by

$$SK_q(g) = \frac{1 - e^{gz_q}}{1 + e^{gz_q}} \tag{15}$$

and

$$KR(g,h) = \frac{(e^{-gz_{.125}} - e^{gz_{.125}})e^{hz^2_{.125}/2} - (e^{-gz_{.375}} - e^{gz_{.375}})e^{hz^2_{.375}/2}}{(e^{-gz_{.25}} - e^{gz_{.25}})e^{hz^2_{.25}/2}}, \tag{16}$$

respectively. The unknown parameter g can be estimated from equation (15) as

$$\hat{g} = \frac{1}{z_q} log_e\Big(\frac{1 - SK_q}{1 + SK_q}\Big)$$

and the parameter h is then estimated by solving the nonlinear equation (16) for h, with g replaced by $\hat{g}$.

3.3 Determine the values of k_l and k_u of a boxplot

We shall now determine the values of k_l and k_u from equation (7) when the distribution of the data set under study is unknown. The PDFs of the generalized distributions discussed in Section 3.2 are usually not available. Even when the PDF is known, the exact expression of the joint PDF $f_{Z_{l:n},Z_{u:n}}(z_{l:n}, z_{u:n})$ of the statistics $Z_{l:n}$ and $Z_{u:n}$ required in (7) would be complicated or intractable. Therefore we shall recourse to determine the values

of k_l and k_u from (7) via an iterative Monte Carlo method summarized in the algorithm below.

The Sample-Mean Monte-Carlo Algorithm

Step 1. Provide an initial starting value of k_l (or k_u), and determine the depths l and u of the lower and upper fourths of a sample of size n.

Step 2. Generate the standardized lower and upper fourths $z_{l:n}$ and $z_{u:n}$ from the lth and uth quantiles of the fitted distribution.

Step 3. To determine k_l, evaluate $I = 1-(1-p_l)^2$ in which $p_l = I_{G_l(y_l)}(1, l-1)$, $y_l = z_{l:n} - k_l(z_{u:n} - z_{l:n})$ and $G_l(y_l) = F(y_l)/F(z_{l:n})$ (to determine k_u, evaluate $I = 1-p_u^2$ in which $p_u = I_{G_u(y_u)}(n-u, 1)$, $y_u = z_{u:n}+k_u(z_{u:n}-z_{l:n})$ and $G_u(y_u) = [F(y_u) - F(z_{u:n})]/[1 - F(z_{u:n})]$).

Step 4. Repeat Step 2 and Step 3 M times to obtain $I_1, I_2, ..., I_M$.

Step 5. Compute $\bar{I} = (I_1+I_2+, ..., +I_M)/M$ and compare $\bar{I}$ with the specified value α and search for a new value of k_l (or k_u) that could yield a value of $\bar{I}$ closer to α by using a univariate direct search method.

Step 6. Repeat Step 2 to Step 5 until the values of $\bar{I}$ converge to α or successive values of k_l (or k_u) converge.

The order statistics $z_{l:n}$ and $z_{u:n}$ of Step 2 are evaluated by adjusting the mean and standard deviation of the fitted distribution, i.e.

$$z_{l:n} = \frac{x_{l:n} - E(X)}{\sqrt{Var(X)}} \text{ and } z_{u:n} = \frac{x_{u:n} - E(X)}{\sqrt{Var(X)}}.$$

The mean, variance, lower and upper fourths of the generalized Tukey's *lambda*-distribution are given by

$$E(X_U) = (\lambda_2 + 1)^{-1} - (\lambda_1 + 1)^{-1},$$

$$Var(X_U) = [(2\lambda_1 + 1)(\lambda_1 + 1)^2]^{-1} + [(2\lambda_2 + 1)(\lambda_2 + 1)^2]^{-1}$$
$$+ 2(\lambda_1\lambda_2)^{-1}[B(\lambda_1 + 1, 1)B(1, \lambda_2 + 1) - B(\lambda_1 + 1, \lambda_2 + 1)]$$

$$x_{l:n} = \frac{u_{l:n}^{\lambda_1} - 1}{\lambda_1} - \frac{(1-u_{l:n})^{\lambda_2} - 1}{\lambda_2},$$
$$x_{u:n} = \frac{u_{u:n}^{\lambda_1} - 1}{\lambda_1} - \frac{(1-u_{u:n})^{\lambda_2} - 1}{\lambda_2},$$

respectively, where $u_{l:n}$ and $u_{u:n}$ are the lower and upper fourths of a pseudo-random sample of size n generated from a uniform (0,1) distribution, which are readily available from, for example, the routine RNUNO of the IMSL library (IMSL, 2006).

The mean, variance, lower and upper fourths of the generalized Tukey's gh-distribution are given by

$$E(X_Z) = \left(g\sqrt{1-h}\right)^{-1}\left(e^{g^2/[2(1-h)]} - 1\right),$$
$$Var(X_Z) = \left(g^2\sqrt{1-2h}\right)^{-1}\left(e^{2g^2/(1-2h)} - 2e^{g^2/[2(1-2h)]} + 1\right) - E^2(X_Z),$$
$$x_{l:n} = [(e^{gz_{l:n}} - 1)/g]e^{hz_{l:n}^2/2},$$
$$x_{u:n} = [(e^{gz_{u:n}} - 1)/g]e^{hz_{u:n}^2/2}$$

where $z_{l:n}$ and $z_{u:n}$ are the lower and upper fourths of a pseudorandom sample of size n generated from a standard normal distribution. Note that the variance of X_Z exists only when $h < 0.5$.

4 Applications

We illustrate the outliers labeling procedure in Section 3 using four data sets taken from a trim-and-form process in an integrated circuit (IC) manufacturing, and a sample of 242 data simulated from a gamma distribution with shape parameter $\nu = 2.0$, scale parameter $\beta = 5.0$ and location parameter $\theta = 30.0$. The robust skewness $SK_{0.125}$ and kurtosis KR, together with the estimated parameters of the fitted generalized Tukey's *lambda*- and gh-distributions of each of these data sets, are given in Table 1. The values of k_l and k_u in the boxplot are then determined from the simulated lower and upper fourths of the fitted distribution using the iterative Monte Carlo

algorithm given in Section 3.3. They are evaluated with outliers error rate $\alpha = 0.05$ and by simulating $M = 500,000$ samples in the Monte Carlo algorithm.

The lead spread measurements of 108 ICs in each of the first four data sets are measured with a high-speed automatic measurement system. The first three data sets are slightly skewed to the right, thus the value of k_u is slightly larger than k_l in these data sets. The first data set has kurtosis value $KR = 1.2$ which is close to that of the normal distribution ($KR = 1.2331$). The second data set has KR=1.0 which leads to shorter tails and thus smaller values of k_l and k_u than those of the first data set and the normal distribution ($k_l = k_u$=2.255, Table 1 of Sim et al., 2005). On the contrary, the third data set has $KR = 1.385$ which leads to longer tails and larger values of k_l and k_u than those of the first data set and the normal distribution. The fourth data set is moderately skewed and its kurtosis value KR is same as that of the first data set. Therefore the values of k_l and k_u of the boxplot, evaluated based on the shape of the fourth data set, are close to their corresponding values of the first data set, however, with the value of k_u larger than k_l.

The first and last parts of each of these four data sets, sorted by increasing values, are

Data set 1: 23, 24, 24, ..., 39, 40, 41, 43, 47

Data set 2: 15, 16, 16, ..., 37, 37, 43, 205, 206

Data set 3: 14, 14, 16, ..., 37, 39, 57, 100, 100, 223, 292

Data set 4: 13, 17, 19, ..., 34, 34, 35, 35, 41

From the lower fence (LF) and upper fence (UF) of the boxplot given in Table 1, we identify the observations in the subsets {47}, {43, 205, 206}, {57, 100, 100, 223, 292} and {13, 41} as the outliers of the data sets 1, 2 , 3 and 4, respectively. The conventional boxplot with $k = 1.5$ over declares the number of outliers in samples 1 and 3, whereas boxplot with $k = 3.0$ under declares the number of outliers in samples 1, 2 and 4.

The fifth data set simulated from a gamma distribution is highly skewed

with $SK_{0.125} = 0.3580$. The first and last parts of the ordered data set are

30.33, 30.51, 30.94, ..., 54.68, 55.53, 56.00, 56.72, 59.11, 59.30,
60.18, 62.30, 62.45, 69.43

Boxplot constructed with the *lambda*-distribution (or *gh*-distribution) fitted to this data set has identified its largest value as an outlier, whereas the conventional boxplot with $k = 1.5$ declares as many as nine data values as outliers.

5 Summary

Graphical boxplot has been widely used for routine screening of outliers in data analysis. The conventional boxplot declares observations in a given data set as outliers if they lie beyond the interval $[Q_1 - k_l(Q_3 - Q_1), Q_3 + k_u(Q_3 - Q_1)]$ with $k_l = k_u = 1.5$ or 3, irrespective of the number of observations and the underlying distribution of the data set. Procedures in constructing boxplot that take into consideration the size and shape of the data set as well as controling the risk of falsely labeled regular observations as outliers are available in the literature, however, only for the case when the underlying distribution of the data set is known. This paper proposed a novel procedure in constructing boxplot to detect outliers in data set taken from unknown underlying distribution. Two generalized Tukey's distributions are used to fit the shape of the given data set based on its robust coefficients of skewness and kurtosis. The lower and upper fences of the boxplot are then determined based on the fitted distribution by using an iterative Monte-Carlo algorithm and an expression given in Sim et al. (2005).

6 References

Banerjee, S., Iglewicz, B. (2007). A simple univariate outlier identification procedure designed for large samples. *Communications in Statistics–*

Simulation and Computation 36:249-263.

Barnett, V., and Lewis, T. (1994). *Outliers in Statistical Data.* 3rd edition. New York: Wiley.

Brys, G., Hubert, M., Struyf, A. (2004). A robust measure of skewness. *Journal of Computational and Graphical Statistics* 13:996-1017.

Carey, V. J., Walters, E. E., Wager, C. G., Rosner, B. A. (1997). Resistant and Test-Based Outlier Rejection: Effect on Gaussian One- and Two-Sample Inference. *Technometrics* 39:320-330.

David, H. A. (1981). *Order Statistics.* 2nd edition. New York: Wiley.

Davies, L., Gather, U. (1993). Identification of Multiple Outliers. *Journal of the American Statistical Association* 88:782-792.

Freimer, M., Kollia, G., Mudholkar, G. S., Lin, C. T. (1988). A study of the generalized Tukey lambda family. *Communications in Statistics–Theory and Methods* 17:3547-3567.

Hawkins, D. M. (1980). *Identification of Outliers.* London: Chapman and Hall.

Hinkley, D. V. (1975). On power transformations to symmetric. *Biometrika* 62:101-111.

Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and-h distributions, in Hoaglin, D. C., Mosteller, F., Tukey, J. W. (eds.), *Exploring Data Tables, Trends and Shapes.* New York: Wiley.

Hoaglin, D. C., Iglewicz, B., Tukey, J. W. (1986). Performance of Some Resistant Rules for Outlier Labeling. *Journal of the American Statistical Association* 81:991-999.

Hoaglin, D. C., Iglewicz, B. (1987). Fine-Tuning Some Resistant Rules for Outlier Labeling. *Journal of the American Statistical Association* 82:1147-1149.

Hoaglin, D. C., Mosteller, F., Tukey, J. W. (2000). *Understanding Robust and Exploratory Data Analysis.* New York: Wiley. [Wiley Classics Library Series].

IMSL (2006). IMSL Fortran Numerical Library: User's Guide. Version 6.0. (http://www.vni.com/products/imsl/documetation/index.php)

Martinez, J., Iglewicz, B. (1984). Some properties of the Tukey g and h family of distributions. *Communications in Statistics-Theory and Methods* 13:353-369.

Moors, J. J. A. (1988). A quantile alternative for kurtosis. *Statistician* 37:25-32.

Sim, C. H., Gan, F. F., Chang, T. C. (2005). Outlier Labeling with Boxplot Procedures. *Journal of the American Statistical Association* 100:642-652.

Tukey, J. W. (1977). *Exploratory Data Analysis.* Reading, MA: Addison Wesley.

Table 1. The values of k_l, k_u, lower fence (LF), and upper fence (UF) of the boxplots constructed with outliers error rate $\alpha = 0.05$, and the distribution fitted based on the robust coefficients of skewness (SK) and kurtosis (KR) of five given data sets. The estimated values of the parameters λ_1 and λ_2 of the generalized *lambda*-distribution, and the estimated values (in *italics*) of the parameters g and h of the generalized gh-distribution are given in column 4.

Data set	Sample Size	Robust skewness & kurtosis	Estimated Parameters $\hat{\lambda}_1$ $\hat{g}$	$\hat{\lambda}_2$ $\hat{h}$	k_l	k_u	LF	UF	Number of outliers: Fitted distribution	$k_l = k_u = k$: $k = 1.5$	$k = 3.0$
1	108	SK =0.111	0.40494	0.10685	2.1042	2.2176	17.4791	44.0881	1	3	0
		KR =1.200	*0.19398*	*-0.05298*	*2.1111*	*2.1906*	*17.4445*	*43.9529*	*1*		
2	108	SK =0.100	1.41433	0.87008	1.6076	1.6492	9.7468	39.5442	3	3	2
		KR =1.000	*0.17444*	*-0.30888*	*1.5921*	*1.6376*	*9.8552*	*39.4630*	3		
3	108	SK =0.077	-0.04580	-0.21192	3.0599	3.1608	1.1106	48.0449	5	6	5
		KR =1.385	*0.13400*	*0.17186*	*3.1230*	*3.2083*	*0.7003*	*48.3537*	5		
4	108	SK =0.250	0.78435	0.02917	2.0894	2.2776	13.5531	40.3879	2	2	0
		KR =1.200	*0.44406*	*-0.10550*	*2.1485*	*2.3248*	*13.2575*	*40.6242*	2		
5	242	SK =0.358	0.96875	-0.14713	2.6446	2.9167	14.1036	65.8923	1	9	1
		KR =1.278	*0.65117*	*-0.08465*	*2.6577*	*2.9007*	*13.9955*	*65.7660*	1		